

# SimLM: Can Language Models Infer Parameters of Physical Systems?

Sean Memery<sup>1</sup> Mirella Lapata<sup>1</sup> Kartic Subr<sup>1</sup>

## Abstract

Several machine learning methods aim to learn or reason about complex physical systems. A common first-step towards reasoning is to infer system parameters from observations of its behavior. In this paper, we investigate the performance of Large Language Models (LLMs) at performing parameter inference in the context of physical systems. Our experiments suggest that they are not inherently suited to this task, even for simple systems. We propose a promising direction of exploration, which involves the use of physical simulators to augment the context of LLMs. We assess and compare the performance of different LLMs on a simple example with and without access to physical simulation.

## 1. Introduction

The rapid evolution of language models has unveiled a powerful form of artificial intelligence that can generate plausible responses to general-purpose queries. These models have been found lacking in some domains and in this paper we demonstrate that they are fundamentally unsuited to reasoning problems involving physical systems. However, due to their sophisticated and articulated interpretation of complex ideas in natural language the hope is that, someday, they will be able to provide useful insights by answering questions such as “Is ‘this’ protein sequence likely to fold into a similar structure to ‘that’ sequence?”.

Indeed, we begin with a far simpler problem and assess the efficacy of LLMs at reasoning about physics—the *inverse* problem. Specifically, we consider predicting the location of the third bounce of a spherical projectile launched from a certain height  $h$  with a certain horizontal initial velocity  $v$ , for different terrain (see Figure 1). This is in contrast to concurrent works that investigate their ability to *predict* behavior (the forward problem).

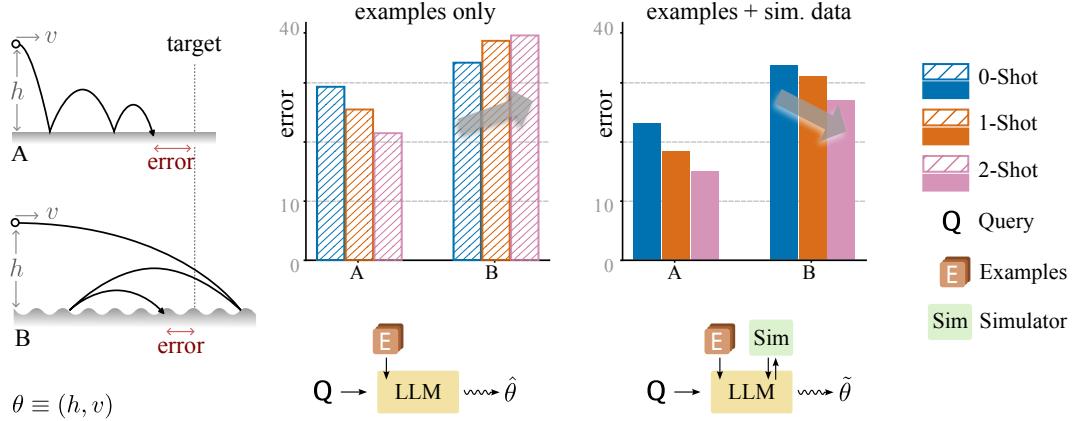
<sup>1</sup>Department of Informatics, the University of Edinburgh, United Kingdom. Correspondence to: Sean Memery <s.memery@ed.ac.uk>.

Many methods (Wei et al., 2022; Kojima et al., 2022) improve reasoning abilities of LLMs via augmentations to their contexts. There have also been methods that rely on external systems to improve the consistency of LLM logic (Liu et al., 2022; Madaan et al., 2022; Lyu et al., 2023). These methods improve capabilities while being independent of the training of the LLM. We follow this paradigm, but focus on physical reasoning. Rather than evaluate a model’s ability to methodically dissect a physics problem into a sequence of solvable problems, as a first step we consider their ability to solve inverse physics problems. Specifically, we assess the abilities of LLMs to *infer initial parameters*  $\theta \equiv (h, v)$  in the above example of projectile motion.

Several methods in machine learning aim to mimic human intuition about complex physical systems (Spelke et al., 1992; Bramley et al., 2018) and develop approximate inference methods (Lake et al., 2016; Tenenbaum, 2018; Ullman et al., 2018) for downstream applications such as *system identification* in robotics (Lopez-Guevara et al., 2017). There is some evidence that LLMs require ‘coaxing’ to elicit reasoning natural to humans (Wei et al., 2022). Few-shot chain-of-thought (CoT), and more sophisticated methods (Yao et al., 2023), are promising directions for natural language, but are they effective for problems involving physics?

In this paper we devise a simple inverse physics problem whose difficulty can be parameterised. Then we compare and analyse the performance of state-of-the-art LLMs at solving this problem. All chosen models perform unsatisfactorily. We find that there is benefit from using chain-of-thought reasoning to augment the context with examples of the forward problem (labeled ‘examples only’ in Figure 1). For a flat surface, task A, the results show that few-shot examples consistently improve performance. However changing the surface to a  $\sin$  wave, task B, adds sufficient complexity to yield diminishing returns.

We propose to ameliorate this by enabling the LLM to query a physical simulator. This is in the spirit of Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), where reference information is retrieved using an external system. In our case, the external system is a physical simulator. Some improvement is obtained via this strategy which we call SimLM (pronounced sim-lim). Despite some success in the simple example, we provide further experiments in



*Figure 1.* We posed a number of Large Language Models a simple query  $Q$ : “With what horizontal velocity  $v$  and from what height  $h$  should a ball be thrown so that its third bounce is within 1 m of 50 m”. Most models are incapable of answering this query. The figure shows the mean errors when the ground is flat (A) as well as sinusoidal (B). Surprisingly, providing more examples worsens the performance when the forward problem is difficult (sinusoidal case). The central proposition in this article is that augmentation of queries/prompts via a physics simulator enhances the ability of LLMs to reason about physics for difficult problems.

Section 7 to draw the attention of the community to this general problem which remains open and exciting.

## 2. Background

### 2.1. Previous Work

**Intuitive physics** There has been some work on the ability of humans to reason about some complex physical systems intuitively without formal training in physics. For example, on the behavior and learning patterns of infants (Spelke et al., 1992; Baillargeon, 1994; Baillargeon et al., 1996) and their ability to learn efficiently from sparse experiences. There are some models of how intuition is built (Battaglia et al., 2013; Lake et al., 2015; Battaglia et al., 2016) surrounding physical events. Recently, generative AI and LLMs (Wu et al., 2015; Duan et al., 2022; Benchekroun et al., 2023) have been used to explore similar ideas. (Jassim et al., 2023) investigate intuitive physical understanding of multi-modal models (language and image understanding) on a data set created from physical simulations towards labeling videos as depicting plausible physics events.

**Prompting and Reasoning** There is considerable interest in devising augmentation and manipulation of the context provided to large language models so that its performance improves for a given task. This general method is known as “prompt engineering” (Liu et al., 2021). Several papers (Nye et al., 2021; Kojima et al., 2022) demonstrate that when an LLM is constrained to solve a task “step by step”, it improves reasoning capabilities by triggering a favorable response for the LLM. When applied to calculating answers to maths word problems, such a dissection of the problem can make the difference between the LLM producing one

incorrect answer to instead answering some sub-problems correctly. This is particularly known to benefit sufficiently large LLMs on tasks involving multi-step computations. With this idea, LLMs have been shown to perform many tasks zero-shot (i.e. no examples given) with appropriate prompting and augmentation. Many approaches have enhanced this fundamental idea recently (Dua et al., 2022; Creswell et al., 2022; Press et al., 2022). (Qiao et al., 2022) provide a thorough survey of the LLM reasoning landscape that catalogs many methods. In particular, it lists many examples of single-step, strategy-enhanced reasoning, i.e. methods that improve reasoning using just a single step addition to the LLM context (Kojima et al., 2022; Zhang et al., 2022; Fu et al., 2022; Zhou et al., 2022). Recently, new directions of LLM reasoning have been introduced that attempt to combine reasoning and action. (Yao et al., 2023) propose a system for iterative reasoning and action with impressive results involving grounded text-based plans within simulated text environments. (Hao et al.) propose a Monte Carlo tree search algorithm to optimize reasoning towards grounded understanding of world models.

**LLMs and Physics** Despite many works that evaluate LLMs on physics data (Xuan-Qui et al., 2023; Ding et al., 2023; Yeadon and Hardy, 2023; Ali-Dib and Menou, 2023), only a handful of them explore the interactions of LLMs with physics environments. For example, (Liu et al., 2022) introduces a system to ground LLM reasoning using physics simulation to answer predetermined physics-based questions. They use a secondary LLM step where the conditions of a physics simulation, defined explicitly in code, and statistics of the outcome are provided in-context to the LLM answering the original question. This yields a significant improvement over standard few-shot methods on physics QA

data. However, it does not perform tasks within a simulated environment. (Benchekroun et al., 2023) introduce a dataset to evaluate the consistency of LLM’s world models, through repeated queries about specific physical situations. A recent benchmark for extensive physical reasoning, called NEWTON (Wang et al., 2023), evaluates the ability of LLMs to predict physical properties of objects. Their method and data highlight the inability of LLMs to predict (forward model) the behavior of physical systems. In this paper we investigate how to improve the ability of LLMs to infer initial conditions of a physical system.

## 2.2. Physical System

The physical system in this paper is simple spherical projectile motion coupled with linear elastic collision at each bounce on the ground, as studied in secondary education level physics. The ground is flat for the simple experiment and bumpy for other experiments. The height  $h_0$  and horizontal velocity  $\vec{v}_0$  are initial conditions and acceleration due to gravity is  $g$ . We assume that the launch-point has an X-coordinate of zero. There may be other forces to consider in a realistic system, such as friction or air resistance but we do not consider them in this paper.

The position  $\vec{x}_t$  of the ball at any given time ( $t$ ) is given by  $\vec{x}_t = \vec{v}_0 t + \vec{g} t^2/2$ . Since the total mechanical energy (kinetic + potential) of an object remains constant when no non-conservative forces are acting upon it (for example air resistance, which is not considered in this work) we have  $mgh_t + mv_t^2/2$  as a constant, where  $m$  is the mass of the object. Finally, when an object collides with the ground, the direction of its new velocity is governed by  $\vec{n}$  the vector normal to the ground at the point of contact. The magnitude of the new velocity depends on the elasticity of the collision quantified by the coefficient of restitution  $e$ . For a perfectly elastic collision,  $e = 1$ , and for a perfectly inelastic collision,  $e = 0$ . In this paper, we use  $e = 0.9$  for all our experiments. The velocity after collision  $\vec{v}'$  is given by  $\vec{v}' = e \vec{r}$ , where  $\vec{r}$  is found by reflecting the incident velocity vector about the normal vector,  $\vec{r} = \vec{v} - 2 \vec{n} (\vec{v} \cdot \vec{n})$ .

## 2.3. Prompting Methods Notation

We extend existing notation (Sumers et al., 2023) that represents LLM prompting methods using production systems. For example,  $X \rightarrow XY$  represents a precondition (input)  $X$  that results in an output of  $XY$ , where the production system specifies the possible values of  $Y$ . This was expanded to include large language models as they can be seen as providing probabilities across productions from which a final production is sampled. This was written as a stochastic production  $Q \rightsquigarrow QA$  where  $Q$  is the query and  $A$  is the LLM generated answer. We adopt this notation as a tool to understand and compare recent prompting methods for

LLMs as well as to denote our contribution. We use different colors to highlight context augmentation inputs (e.g.  $O$ ) vs LLM-generated outputs (e.g.  $A$ ).

**Retrieval-Augmented Generation** As LLMs are inherently stateless, retrieval is a method commonly used to provide a working memory for the model. It can also be used to retrieve specific information that the LLM may use to improve on its completion. For example, a simple retrieval of relevant Wikipedia pages would improve an LLM’s question-answering capability. This method is known as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and can be denoted as  $Q \rightarrow QO \rightsquigarrow QOA$  where  $O$  is the retrieved information.

**Self-Critique** (Saunders et al., 2022) introduced self-critique as a simple method that asks the LLM to critique its own output. For question-answering this is as simple as the LLM answering the question, critiquing its own answer, and answering again. Using the same notation as above, this can be written as a sequence of stochastic productions

$$Q \rightsquigarrow QA_0 \rightsquigarrow QA_0C \rightsquigarrow QA_0CA_1 \quad (1)$$

where  $C$  is the self-critique generated by the LLM based on its own initial answer  $A_0$  which is then updated to  $A_1$ .

**Few-shot Chain-of-Thought** (Wei et al., 2022) introduced chain-of-thought (CoT) prompting, a method to improve the reasoning abilities of LLMs. CoT prompting provides the LLM with in-context examples of question-reasoning-answer sequences. More of these examples has been shown to lead to better performance in general. This is visualized below, with each example  $E$  containing full chain-of-thought:

$$Q \rightsquigarrow E_1 \dots E_N QR \rightsquigarrow E_1 \dots E_N QRA \quad (2)$$

## 3. SimLM

We now describe our method SimLM of using a physical simulator in conjunction with existing work on prompting.

### 3.1. Method

Our prompting method inserts physics simulations between reasoning and critiquing steps. It can be viewed as a retrieval augmented generation combined with self-critique  $C$  of its reasoning—here the “retrieve” operation corresponds to physics simulation feedback  $P$ . Thus we can write this as a sequence of four productions where the second production is the execution of a physics simulator:

$$\begin{aligned} Q &\rightsquigarrow QR \\ &\rightarrow QRCP \\ &\rightsquigarrow QRPC \\ &\rightsquigarrow QRPCA \end{aligned} \quad (3)$$

We apply this iteratively  $N$  times, so that the model is allowed to improve on its reasoning via multiple physics simulations and self-critique steps i.e. generate reasoning  $R$ , augment with feedback  $P$ , generate self-critique  $C$  and repeat  $N$  times before producing the final answer  $A$ :

$$\begin{aligned} Q &\rightsquigarrow QR \\ &\rightsquigarrow QR P_1 C_1 \\ &\rightsquigarrow QR P_1 C_1 \dots P_n C_n \\ &\rightsquigarrow QR P_1 C_1 \dots P_n C_n A \end{aligned} \quad (4)$$

We refer to this method as ‘zero-shot with physics’.

Providing in-context examples (Brown et al., 2020) to the LLM improves performance on many tasks. Chain-of-thought prompting showed this to be true for reasoning. Inspired by these findings, we extend our zero-shot method (Equation 4) to include examples  $E_i$ ,  $i = 1, 2, 3, \dots, m$ . Assembling these elements, our full method is:

$$\begin{aligned} Q &\rightarrow E_1 \dots E_m Q \\ &\rightsquigarrow E_1 \dots E_m Q R \\ &\rightsquigarrow E_1 \dots E_m Q R P_1 C_1 \\ &\rightsquigarrow E_1 \dots E_m Q R P_1 C_1 \dots P_n C_n \\ &\rightsquigarrow E_1 \dots E_m Q R P_1 C_1 \dots P_n C_n A \end{aligned} \quad (5)$$

This can also be summarized succinctly as

$$Q \rightsquigarrow \{E\}^* QR \rightsquigarrow \{E\}^* QR \{PC\}^+ A \quad (6)$$

where we use  $*$  and  $^+$  as used in regular expressions to mean any number of instances and at least one instance respectively.

### 3.2. Implementation

**Physics Simulation** To perform the physics simulation, we use the `pymunk`<sup>1</sup> package. This is a lightweight physics package for Python that is built on top of the `Chipmunk`<sup>2</sup> 2D physics simulator. Pymunk performs rigid-body physics simulation with high efficiency and accuracy. To reduce errors caused by the physical limitations of the simulation, the frames per seconds of the simulation is set to 1000Hz. Doing so limits precision errors during physics calculation causing the simulation to be as deterministic as possible.

Using Pymunk, any arbitrary function  $f(x)$  can be specified as the ground geometry by defining  $x$  coordinate limits, e.g.  $[-1000, 1000]$ , and iterating between these limits with some step  $c$ . Then, a list of  $y$  coordinates can be calculated as  $y = f(x)$ , and used with the  $x$  positions to create a piecewise linear approximation to the function.

**LLM Inference** We perform LLM inference using the `guidance`<sup>3</sup> library. `guidance` allows for consistent,

template based prompting of multiple models, local and API-based. Using this, we created a pipeline to prompt LLMs, run simulations, show simulation results to those LLMs, and continue prompting. Algorithm 1 gives an overview of this pipeline. See the supplementary information for more detailed examples and prompts.

---

#### Algorithm 1 LLM prompting in SimLM

---

$$\begin{aligned} Ex &= \text{retrieve}(Q) \\ R, A &= \text{LLM}(Ex, Q) \\ P_1 &= \text{sim}(A) \\ C_1, A_1 &= \text{LLM}(Ex, Q, R, P_1) \\ P_2 &= \text{sim}(A_1) \\ \dots \\ C_N, A_N &= \text{LLM}(Ex, Q, \dots, P_N) \end{aligned}$$


---

An important aspect of our method is the number of attempts of prediction and self-critique. In our implementation, we prompt the LLM to finish repeating attempts if it finds that it is meeting the requirements of the task. We found a suitable limit to be  $N = 5$  and use it in all experiments.

Many works have shown the importance of supplying LLMs with high quality examples for the target task. In our method, the LLM is given examples to act as long-term experience gain by the model. We use previous successful runs of an experiment as examples for future runs. In each experiment we have a success criteria, if this is met then the entire context is saved in a database of reasoning examples. In order to limit the context size of future trials, we then perform basic summarisation on the saved examples. See the supplementary information for summarisation details. Then, a database of summarized examples can be sampled from to provide future trials with in-context CoT examples. In order to prevent performance leakage, we limit the retrieval to examples created by *the same model on the same experiment*.

## 4. Experiments

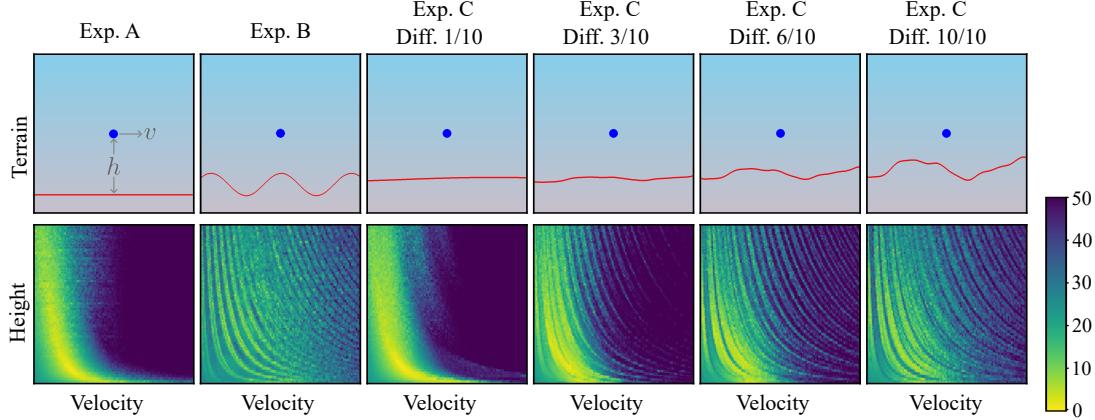
We conducted three experiments (A, B, and C) where LLMs are asked to predict the initial conditions of a ball, such that its third bounce will occur within 1m of a target distance of 50m. The target and tolerance is fixed across all experiments for consistency. Across the variants we vary only the shape of the terrain the ball bounces upon. The top row in Figure 2 illustrates a few instances of this problem showing the ball (blue dot) and the terrain (red line).

The LLMs output two values, as a JSON object: `height` and `horizontal_velocity` which are used to initialize the ball in the simulation. Other initial conditions, such as starting  $x$  position and vertical velocity, are set to zero. We compare our method (SimLM) against a baseline Chain of Thought (CoT) method. The baseline simply uses the first set of initial conditions output by the model, with no

<sup>1</sup><https://github.com/viblo/pymunk>

<sup>2</sup><https://chipmunk-physics.net/>

<sup>3</sup><https://github.com/guidance-ai/guidance>



**Figure 2.** Visualizations capturing the difficulty of problems as a function of the ground surface (top row) and mean error across trials (bottom row). Errors (heat map colors) are shown across the parameter space spanned by horizontal velocity (X-axis) and initial height (Y-axis) measured as the mean distance away from the target (50m).

physics simulation feedback. This “blind” attempt at reasoning about the initial conditions does not update itself based on simulation results. Our method uses feedback from simulation results and the LLM’s self-critique.

We provide examples to the LLM that are gathered from previous runs of the same experiment. This is consistent across all experiments, but retrieved examples are not using the same target value as the current trial. We retrieve a random selection of examples that are sampled uniformly across targets, with buckets of 10m. Thus each trial is equally likely to contain examples that are near or far from its target (rather than biased examples around the target).

**Flat Surface** Experiment A is a simple baseline of the performance of LLMs with and without our method. The surface that the ball is bouncing on is a flat plane described by  $y = 0$ . There are still however multiple unknown factors such as the coefficient of restitution of the ball and surface to consider. This makes it difficult for an LLM to explicitly calculate the required parameters, as we want to evaluate the LLM’s common sense physical reasoning rather than its computation abilities. Column 1 of figure 2 shows a visualization of this experiment.

**Uneven Surface** Experiment B is considered a non-linear step from experiment A. On a flat surface, each bounce should be a further distance from the ball’s origin. However in experiment two, the surface is described by the sinusoid  $y = \alpha \cdot \sin(f \cdot x)$ , where  $\alpha$  is the amplitude of the sinusoid, and  $f$  is the frequency. This can lead to the ball colliding with the surface at an incline and bouncing backwards, causing the ball’s bounce positions to no longer be monotonic. For an LLM, this is a particularly difficult challenge as examples are not as helpful when a small difference in position

can lead to extremely different outcomes. The LLM must then rely heavily on particularly relevant examples and trial-and-error to improve its prediction. We fix the amplitude and frequency of the sinusoid to 1. Column 2 of figure 2 shows a visualization of this experiment.

**Varying Difficulty** In experiment C we explore how the performance of an LLM degrades as a surface becomes more “difficult”—that is more chaotic or rough. The parameter space heatmaps of figure 2 illustrate the difficulty posed to an optimizer during inference across experiments. For different terrains, as the area in the heatmap where the error is lowest becomes sparser and less defined, we classify the terrain as becoming more “difficult” to predict. We generate a range of settings by interpolating between two surfaces, one considered easy and one considered difficult. The former is a single sinusoid  $y_E = 0.15 \cdot \sin(0.25x)$  while the latter is defined as a sum of sinusoids:

$$y_H = 0.6 \cdot \sin(0.9x) + 0.15 \cdot \sin(2.25x) + 0.05 \cdot \sin(4.5x)$$

Then, to create a surface with a difficulty rating of  $d \in [0, 1]$ , a simple linear interpolation is performed between the two surface functions. We create 10 different surfaces for experiment three, of uniformly spaced difficulty, in order to evaluate how baseline CoT and SimLM adapt to increasingly difficult surfaces. Columns 3 – 6 of figure 2 visualize these surfaces, for difficulties 1/10, 3/10, 6/10, and 10/10.

## 5. Results

We tabulate mean errors obtained using different models (Table 1) across Experiments A and B. We calculate an average error, in each case, across  $\sim 1000$  instances. We

		Baseline CoT			SimLM		
		0-Shot	1-Shot	2-Shot	0-Shot	1-Shot	2-Shot
PaLM-2	Exp. A	34.84	28.51	22.41	24.21	19.34	15.03
	Exp. B	32.12	35.53	37.54	31.70	30.72	27.05
GPT-3.5-turbo	Exp. A	19.23	18.68	16.40	20.18	16.34	15.41
	Exp. B	37.03	40.39	40.19	42.17	36.20	29.10
Llama-2-70b-chat	Exp. A	29.14	25.95	21.67	23.64	20.53	19.45
	Exp. B	40.91	41.62	40.32	40.12	35.54	30.15
Llama-2-13b-chat	Exp. A	30.18	26.48	22.96	26.52	23.53	20.87
	Exp. B	43.89	44.68	43.63	42.19	39.67	31.16
Llama-2-7b-chat	Exp. A	41.95	29.03	27.99	42.73	29.38	25.16
	Exp. B	43.02	41.09	44.02	42.10	34.87	31.30

Table 1. The average absolute error (for a target of 50m) of LLMs for experiments A and B, for baseline CoT and SimLM methods.

plot<sup>4</sup> comparisons for the models PaLM-2 (chat-bison variant) and GPT-3.5-turbo (figure 4).

The results of experiment 3 for the PaLM-2 model are presented in figure 3. Here, surface difficulties are grouped into the categories *easy* (difficulty 1-3), *medium* (difficulty 4-7), and *hard* (difficulty 8-10). For each category, the results of the corresponding surfaces are averaged and displayed as a relative error of SimLM over baseline CoT. That is, a relative error value below 1 signifies better performance of SimLM and decreasing relative error means a growing disparity between the SimLM results and baseline CoT.

We only include average performances and selected statistics in the paper. Complete results of each experiment are included in the appendix. For each comparison between baseline and non-baseline results, two-sample t-tests were performed to assess whether there were statistically significant differences in their distributions. In each test, the null hypothesis, that there is no difference between the groups, was tested. A threshold p-value of 0.05 was used to determine statistical significance. In each case the p-value was below this threshold, and the null hypothesis was rejected, suggesting that the differences observed in the results were unlikely to have occurred by chance.

## 6. Discussion

### 6.1. Models Compared

We focus on three classes of models as a selection of currently feasible models (as of December 2023), based on a balance of efficiency, cost, and capability:

- GPT-3.5-turbo: The API accessible chat model from OpenAI (Brown et al., 2020).
- PaLM-2: The API accessible model from Google (Anil et al., 2023), we used the chat-bison variant.

<sup>4</sup>The colors chosen for plots were based on (Wong, 2011), to maximise readability for colour blindness.

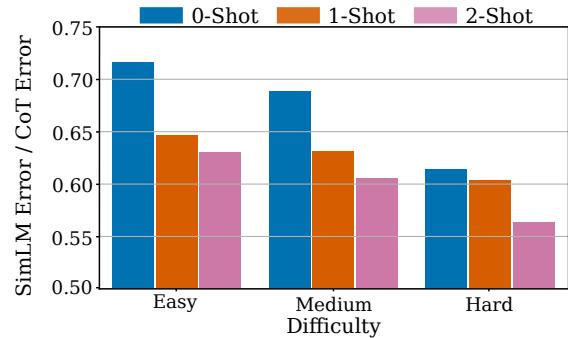


Figure 3. A plot of the error ratio of our method vs baseline for surfaces with varying difficulty. The error ratio is less than one showing that SimLM outperforms the baseline. As the problem becomes more difficult, SimLM is relatively more effective.

- Llama-2: Open source, local chat models from Meta (Touvron et al., 2023) with 7b, 13b, and 70b parameters.

### 6.2. Simulation Improves Performance

In figure 4, we observe little difference between the LLM performance across methods in experiment A. However, notice the decrease in error between experiment B with no physics feedback (baseline CoT) and experiment B with physics feedback (SimLM). We see a small drop in error for SimLM over the baseline, but significantly the baseline CoT does not effectively make use of examples in experiment B, and performance actually suffers as more examples are used. On the other hand, SimLM uses simulation to enable effective use of examples and improve performance. This is an important aspect of our method, to improve physical reasoning on uneven terrains.

### 6.3. Relative Error Decreases as Difficulty Increases

Figure 3 highlights the performance increase of SimLM over the baseline for the best performing model, PaLM-2 (chat-bison variant). As the difficulty of the surface increases the overall relative error decreases. SimLM is outperforming baseline CoT more as the surface becomes more difficult and uneven. We contribute this to the LLM relying more on the feedback received from the physics simulation, as the surface becomes more difficult to predict. We also see that the relative error of the two methods is below 1 in all cases, showing the strength of SimLM on uneven surfaces generally, and that the relative error decreases as more examples are used in each method, meaning SimLM is benefiting more from previous examples than baseline CoT.

### 6.4. Large Models Outperform Small

As a further study of the capabilities of smaller models, in particular the popular range of Llama-2 models, we repeat experiments A and B for the Llama-2 models of parameter sizes 7b and 13b, and compare these results to the largest Llama-2 model of parameter size 70b. Table 1 shows the results of the model size study.

We observe a similar trend that is seen in the larger, API based, models. Experiment A is still a relatively achievable task for the models, but as the difficulty is increased in experiment B the same loss of CoT improvements occur. The overall difference between models is simply a decreasing average error as the model size increases, as generally expected. This is a sign that larger models perform better at physical reasoning than their smaller counterparts, meaning there is an ability inherent in the models that is being improved on when parameter counts are scaled up.

## 7. Generalisation and Future Work

Although our proposed idea was useful in inferring two parameters of interest, we encountered difficulties in higher-dimensional settings. For example, it was not very successful on an experiment, related to the physics on a pool/billiards table, with five parameters. Pool has been used for research purposes (Leckie and Greenspan, 2006; Smith, 2007). We explain the experiment and have set up an evaluation framework (which we will share) with leaderboards in the hope that evaluation of future work in this direction may be streamlined.

We posed LLMs with the query ‘‘Pot the red ball into the back left pocket’’ and measured their success rate (see Table 2). In order to provide the LLM with a physically accurate 3D environment, we made use of the PoolTool<sup>5</sup>

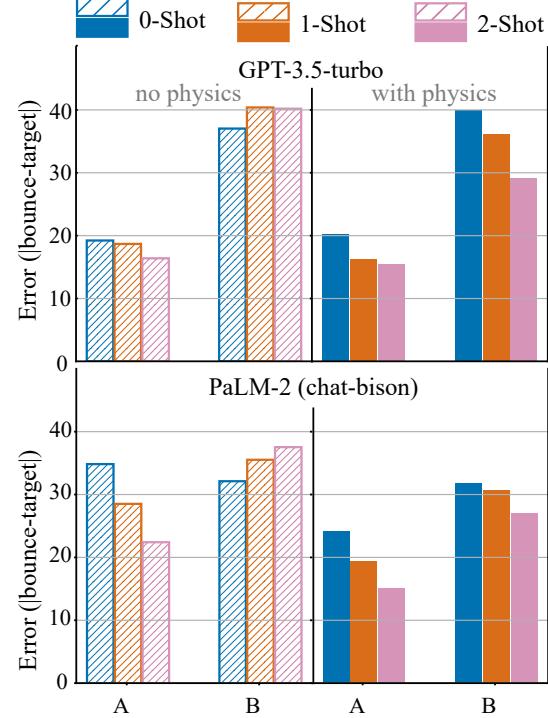


Figure 4. The use of physics simulation data helps LLMs (GPT and PaLM) cope with increasing difficulty. Plots of error in choosing initial conditions for a ball to bounce at a target distance on a flat surface (experiment A) and on an uneven surface (experiment B).



Figure 5. An LLM taking a shot using the *PoolTool* simulator.

(Random = 0.4%)	CoT Baseline	SimLM
GPT-3.5-turbo	0.7%	1.1%
PaLM-2	1.4%	2.0%
Llama-2-13b-chat	0.6%	0.9%
Llama-2-7b-chat	0.5%	0.8%

Table 2. Success rate of LLMs potting balls in 3D pool.

<sup>5</sup><https://github.com/ekiefl/pooltool>

application, a highly accurate simulation of the game of pool. `PoolTool` is capable of simulating ball spin, sequences of ball collisions, cushion reflections, and many other physical interactions relevant to pool/billiards. In order to take a shot in `PoolTool`, five parameters must be set:

1.  $v$ : The speed at which the cue ball will travel,  $v \sim [0, 5]$ .
2.  $\theta$ : The angle of the cue, in order to impart different levels of spin,  $\theta \sim [0, 90]$ .
3.  $\phi$ : The azimuth of the shot, the angle corresponding to the direction the cue ball will travel in,  $\phi \sim [0, 360]$ .
4.  $a$ : The east-to-west coordinate of the point of contact of the cue to the cue ball,  $a \sim [-1, 1]$ .
5.  $b$ : The north-to-south coordinate of the point of contact of the cue to the cue ball,  $b \sim [-1, 1]$ .

Through the combination of these parameters, many different kinds of shots can be taken. For example, in order to apply backspin to the cue ball  $\theta$  should be set to a high value and/or  $b$  should be set to a low value.

We extended `PoolTool` to include string representations of events in the simulation of each shot. A sequence of events can then be represented as a chain of sentence descriptions, for example:

- `CollisionEvent (b1, b2) → "Ball b1 collided with ball b2 at position (x,y)."`
- `PotEvent (b, p) → "Ball b fell into pocket p."`

We also include a textual description of the current board state, before and after each shot using a chain of "`ball_id=(x, y)`" for each ball with position  $(x, y)$  on the table. The positions of the pockets are also given. All of this information is included when prompting the LLM, as we did in our original 2D problem. However, the information needed to describe a single simulation run consists of a large body of text. In order to remain in the context limit of most models, the method was allowed at most  $N = 3$  retries at the task before resulting in failure if there is no success. We also fix the examples used for baseline and SimLM to 3, and retrieve previously successful attempts similarly to the 2D case.

While a wider variety of tasks is supported, for this study we limit tasks to the most basic action in pool: *potting a single ball*. We task LLMs with potting a ball (placed randomly within 10cm from the table's center) into each pocket on the table, and verify the results by finding the correct event in the event log of the simulation. Our results are shown in table 2. The pass rate shown is the percentage of successful trials from  $\sim 1000$  total trials of each method and model pairing. These results come to a very different outcome than the 2D case. Here, the addition of feedback from the simulation improves performance only marginally. We find that across LLMs and methods, there are very few successful trials, and these cases often contain flawed or unsatisfactory logic for the problem. We find that all the

LLMs tested perform quite poorly at this type of task, independent of size. With larger LLMs, it has been shown that performance follows size for many metrics, but this hits a limit for complicated 3D reasoning task such as this, where no matter the quality of reasoning, no LLMs are currently capable of a reasonable pass rate. While still performing better than random selection, the LLMs are attempting the task but simply are not currently capable enough to succeed.

## 8. Conclusion

In this study, we have highlighted the poor performance of LLMs at parameter inference for physical reasoning. We proposed the idea of exposing LLMs to simulation data to improve their performance. Our first attempt is to mimic human learning through experience (trial) and error, which appears to improve performance in 2D. Our method outperforms baseline CoT in 2D environments, and gains further improvements as the difficulty of the environment increases. However our work suggests that additional methods are needed for LLMs to effectively perform actions in simulated environments. SimLM is a positive step towards effective reasoning within complex environments, but when pushed to a full 3D environment those improvements are lost, showing the upper limit of scene complexity that current LLMs can reasonably interact with.

We believe this study has presented evidence of a fundamental shortcoming of LLM capability. Although this outcome may be expected since they are trained on text alone, it is important to highlight such weaknesses considering their massive popularity. This could be useful information to approaches that seek to utilise LLMs in real/simulated environments because it shows that these models have poor/no intuition about physical reasoning.

A natural question from the reader might be “why LLMs?”. While it is evident that several machine learning techniques have been developed particularly for parameter inference and system identification, the hope here is that the scalability of LLMs will create opportunities for further reasoning with natural language about the system. As these models evolve, it is foreseeable that they will offer more capability to reason about physics at a higher level. For example to answer a query such as: “Is it possible to pot the 7-ball without displacing any other ball on the table?” with reference to the state visualized in Figure 5.

## References

- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916, 2022.
- Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M. Dai. Mind’s Eye: Grounded Language Model Reasoning through Simulation, 2022.
- Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. Language models of code are few-shot commonsense learners. *ArXiv*, abs/2210.07128, 2022.
- Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. *ArXiv*, abs/2301.13379, 2023.
- Elizabeth S. Spelke, Karen Breinlinger, Janet Ann Whitson Macomber, and Kristen C. Jacobson. Origins of knowledge. *Psychological review*, 99 4:605–32, 1992.
- Neil R. Bramley, Tobias Gerstenberg, Joshua B. Tenenbaum, and Todd M. Gureckis. Intuitive experimentation in the physical world. *Cognitive Psychology*, 105:9–38, 2018.
- Brenden M. Lake, Tomer David Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. What can the brain teach us about building artificial intelligence? *Behavioral and Brain Sciences*, 40, 2016.
- Joshua B. Tenenbaum. Building machines that learn and think like people. In *Adaptive Agents and Multi-Agent Systems*, 2018.
- Tomer David Ullman, Andreas Stuhlmüller, Noah D. Goodman, and Joshua B. Tenenbaum. Learning physical parameters from dynamic scenes. *Cognitive Psychology*, 104:57–82, 2018.
- Tatiana Lopez-Guevara, Nick K. Taylor, Michael U Gutmann, Subramanian Ramamoorthy, and Kartic Subr. Adaptable pouring: Teaching robots not to spill using fast but approximate fluid simulation. In *Conference on Robot Learning*, 2017.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models, 2023.
- Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401, 2020.
- Renée Baillargeon. How do infants learn about the physical world? *Current Directions in Psychological Science*, 3 (5):133–140, 1994. doi:[10.1111/1467-8721.ep10770614](https://doi.org/10.1111/1467-8721.ep10770614).
- Renee Baillargeon, Laura Kotovsky, and Amy Needham. The acquisition of physical knowledge in infancy. In *Causal Cognition: A Multidisciplinary Debate*. Oxford University Press, 07 1996. ISBN 9780198524021. doi:[10.1093/acprof:oso/9780198524021.003.0004](https://doi.org/10.1093/acprof:oso/9780198524021.003.0004).
- Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110:18327 – 18332, 2013.
- Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350:1332 – 1338, 2015.
- Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Neural Information Processing Systems*, 2016.
- Jiajun Wu, Ilker Yildirim, Joseph J. Lim, Bill Freeman, and Joshua B. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Neural Information Processing Systems*, 2015.
- Jiafei Duan, Arijit Dasgupta, Jason Fischer, and Cheston Tan. A survey on machine learning approaches for modelling intuitive physics, 2022.
- Youssef Benchekroun, Megi Dervishi, Mark Ibrahim, Jean-Baptiste Gaya, Xavier Martinet, Grégoire Mialon, Thomas Scialom, Emmanuel Dupoux, Dieuwke Hupkes, and Pascal Vincent. Worldsense: A synthetic benchmark for grounded reasoning in large language models. *ArXiv*, abs/2311.15930, 2023.
- Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. 2023.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hi-roaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55:1 – 35, 2021.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models, 2021.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. Successive prompting for decomposing complex questions. In *Conference on Empirical Methods in Natural Language Processing*, 2022.

Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. *ArXiv*, abs/2205.09712, 2022.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *ArXiv*, abs/2210.03350, 2022.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. Reasoning with language model prompting: A survey. *ArXiv*, abs/2212.09597, 2022.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models, 2022.

Yao Fu, Hao-Chun Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. *ArXiv*, abs/2210.00720, 2022.

Hattie Zhou, Azade Nova, H. Larochelle, Aaron C. Courville, Behnam Neyshabur, and Hanie Sedghi. Teaching algorithmic reasoning via in-context learning. *ArXiv*, abs/2211.09066, 2022.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with Language Model is Planning with World Model.

Dao Xuan-Quy, Le Ngoc-Bich, Phan Xuan-Dung, Ngo Bac-Bien, and Vo The-Duy. Evaluation of chatgpt and microsoft bing ai chat performances on physics exams of vietnamese national high school graduation examination. 2023.

Jingzhe Ding, Yan Cen, and Xinyuan Wei. Using large language model to solve and explain physics word problems approaching human level. *ArXiv*, abs/2309.08182, 2023.

Will Yeadon and Tom Hardy. The impact of ai in physics education: A comprehensive review from gcse to university levels. 2023.

Mohamad Ali-Dib and Kristen Menou. Physics simulation capabilities of llms, 2023.

Yi Ru Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivas. Newton: Are large language models capable of physical reasoning?, 2023.

Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents, 2023.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

Bang Wong. Points of view: Color blindness. *Nature methods*, 8:441, 06 2011. doi:[10.1038/nmeth.1618](https://doi.org/10.1038/nmeth.1618).

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

William H. Leckie and Michael A. Greenspan. An event-based pool physics simulator. In *Advances in Computer Games*, 2006.

Michael Smith. Pickpocket: A computer billiards shark. *Artif. Intell.*, 171:1069–1091, 2007.